Case Studies on Mutually Enriched Data Projects

Internatioal Data Week – SciDataCon 2018
Gaborone, Botswana

Tyng-Ruey Chuang
with Huang-Sin Syu and Wen-Ting Yang

Institute of Information Science
Academia Sinica
Taipei, Taiwan

Case Studies

- MapBox – Bing Maps Imagery – Strava – OSM
- ClinGen/GlinVar – Hypothes.is – RRID
  - informed by Dr. Chunnan Hsu
- depositar.io – CKAN – Wikidata
- Dadu Plateau Investigation Project
  - presented by Ilya Lee

- **Project descriptions liberally taken from Wikipedia**

Bing Maps Imagery

“Bing maps frequently update and expand the geographic areas covered by their imagery, with new updates being released on roughly a monthly basis. Each imagery release typically contains more than 10TB of imagery.

However, the necessary time-lapse before images are updated means that aerial and Bird's-Eye images for a particular location can sometimes be several years out-of-date.”

Strava

“Strava is a social fitness network that is used to track cycling, running, and swimming activities, among others, using GPS data. Activities are recorded via the Strava mobile application or GPS-enabled fitness watches or cycling computers. Users can upload activities to the Strava site directly, via the Strava mobile application, or via one of Strava's data partners...”

OpenStreetMap

“OpenStreetMap (OSM) is a collaborative project to create a free editable map of the world. Rather than the map itself, the data generated by the project is considered its primary output. The creation and growth of OSM has been motivated by restrictions on use or availability of map information across much of the world, and the advent of inexpensive portable satellite navigation devices. OSM is considered a prominent example of volunteered geographic information. …”

MapBox

“Mapbox is a large provider of custom online maps for websites and applications… Since 2010, it has rapidly expanded the niche of custom maps… Mapbox is the creator of, or a significant contributor to some open source mapping libraries and applications, including the MBTiles specification, the TileMill cartography IDE, the Leaflet JavaScript library, and the CartoCSS map styling language and parser.”

MapBox – Bing Maps Imagery – Strava – OSM

- **“Data Misalignment issues in Taiwan” (2016-09-13)**

  “… Mapbox Data Team has been reviewing the OpenStreetMap Data quality in Taiwan with respect to Satellite imagery. By running a comparison against Bing Imagery and available GPS traces like Strava Heat Map we were seeing a notable data offset of 15-20 meters with respect to GPS Traces.”
  https://www.openstreetmap.org/user/srividya_c/diary/39482

- **“Mapping and Aligning road network in Taiwan based on new Mapbox imagery and Strava” (2016-11-07)**

  “Over the past 7 weeks, Mapbox Data team completely concentrated on aligning and improving the road network of Taiwan. With regular interaction with the community in our mapping ticket and a series of diary posts about our findings, progress and learnings, we have reached the finish line of the task.”

  “Using updated Mapbox satellite imagery and Strava heat map and different offsets with respect to Bing imagery, the data team has aligned 64,680kms of road network in Taiwan which forms the 54.4% of entire Taiwan road network. The overall project consisted of 43 tasking manager tasks covering 27,000 sq kms which is 74% of total area of the country.”
  https://www.openstreetmap.org/user/srividya_c/diary/39839

ClinGen and ClinVar Partnership

- “ClinVar and ClinGen, two NIH-based efforts, have formed a critical partnership to improve our knowledge of clinically relevant genomic variation. This partnership includes significant efforts in data sharing, data archiving, and collaborative curation to characterize and disseminate the clinical relevance of genomic variation.”
- “ClinVar is an archival database that aggregates information about genomic variation and its relationship to human health.”
- “The Clinical Genome Resource (ClinGen) … aims to create an authoritative central resource that defines the clinical relevance of genes and variants for use in precision medicine and research.”

https://www.clinicalgenome.org/about/about-the-clingen-and-clinvar-partnership/

The ClinGen and ClinVar Partnership

Both provide resources to support genomic interpretation

ClinGen - A Program

An NIH funded project

Building a central resource that defines the clinical relevance of genes and variants

ClinGen is addressing the following critical questions:

- Is the gene associated with disease?
- Is the variant pathogenic?
- Is the variant/gene information actionable?

Encouraging **data sharing**

- Promote lab submissions to ClinVar
- Facilitate patient data sharing through GenomeConnect

ClinGen
Clinical Genome Resource

Assessing the clinical **validity** and **actionability** of genes and their relationship to diseases

Expertly **curating** and **interpreting** variants

- Provide curated knowledge to **ClinVar** and on **clinicalgenome.org**

Partnership to improve knowledge of genomic variation

Expert Curation

ClinVar- A Database

- Funded by intramural NIH funding
- Freely accessible and downloadable public archive of reports of the relationship between variants and conditions
- Maintained by the National Center for Biotechnology Information (NCBI)

NCBI
ClinVar

Supporting **sharing** of variants interpretations

Maintaining a publicly available **database** of:

- Interpretations of the clinical significance of variants
- Submitter information
- Supporting evidence and individual level data, when available

Find out more online...

ClinGen

https://www.clinicalgenome.org/
@clingenresource
ClinGen Resource
ClinGen Youtube Channel

ClinVar

http://www.ncbi.nlm.nih.gov/clinvar/
@NCBI_Clinical

Hypothes.is

“Hypothes.is is a 501(c) open-source software project that aims to collect comments about statements made in any web-accessible content, and filter and rank those comments to assess each statement's credibility. It has been summarized as 'a peer review layer for the entire Internet.'

As of December 2017, Hypothes.is is the 113,770th most popular site on the Internet according to Alexa.com.”

RRID

“The Resource Identification Initiative (#RRID) is designed to help researchers sufficiently cite the key resources used to produce the scientific findings reported in the biomedical literature.

Resources (e.g. antibodies, model organisms, and software projects) reported in the biomedical literature often lack sufficient detail to enable reproducibility or reuse. For example, catalog numbers for antibody reagents are infrequently reported, and the version numbers for software programs used for data analysis are often omitted. This has been called out as a serious enough problem by the NIH to introduce new guidelines for Rigor and Transparency for almost all awards in starting in May of 2016.

These guidelines argue for authentication of key research resources, and transparency of how they are reported.”

https://www.force11.org/group/resource-identification-initiative

ClinGen/GlinVar – Hypothes.is – RRID

Dr. Chunnan Hsu (UCSD Medical School):

“Modern biomedical sciences are data-driven, and depend on many reliable databases of biomedical knowledge, or simply 'knowledge bases'. … These knowledge bases contain known associations between genetic variants, disease conditions and/or treatments.

In this talk, I will … propose a potential solution — a hybrid human-AI solution with web annotation and its use by ClinGen (curation for ClinVar) and other biomedical knowledge bases.”

http://www.iis.sinica.edu.tw/HTML/seminar/DJ180062_en.html

CKAN

“The Comprehensive Knowledge Archive Network (CKAN) is a web-based open source management system for the storage and distribution of open data. Being initially inspired by the package management capabilities of Debian Linux, CKAN has developed into a powerful data catalogue system that is mainly used by public institutions seeking to share their data with the general public.

CKAN's codebase is maintained by Open Knowledge International. The system is used both as a public platform on Datahub and in various government data catalogues, …”

Wikidata

“Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation. It is intended to provide a common source of data which can be used by Wikimedia projects such as Wikipedia, and by anyone else, under a public domain license. This is similar to the way Wikimedia Commons provides storage for media files and access to those files for all Wikimedia projects, and which are also freely available for reuse. Wikidata is powered by the software Wikibase.”

depositar.io

“The site data.depositar.io is built by the researchers for the researchers. You are free to deposit, discover, and reuse datasets on depositar for all your research purposes.

Datasets — collections of tables, maps, documents, or other data files — are fast to upload and preview. Once uploaded, datasets have permanent links for all to reference and download. …

Project *depositar* is a work in progress. Both source code and user manual are available.”

https://data.depositar.io/en/about

depositar.io – CKAN – Wikidata

“Moreover, in depositar all datasets use keywords sourced from Wikidata! Datasets on depositar are inter-connected by their semantics. …

The software for depositar is built on top of CKAN, an open source software package for publishing open data. … We customize and extend the CKAN codebase to better support research data management. Our extensions to CKAN are open source too.”

https://data.depositar.io/en/about

Finding Motifs of Mutual Enrichment

- Improving data quality by validation with multiple data sources
- Bootstrapping project development by grounding on popular data resources
- Getting broader participation to existing projects by establishing new data relationships
- Building on top of free software and open data
- Demand-driven cross-project fertilization
- Citizens + task management + automation